DOCUMENT RESUME

ED 181 077 TM 010 038

AUTHOR Lockheed, Marlaine E.
TITLE Sex Bias in Educational Aptitude and Achievement Tests: What Do We Know? Draft.
PUB DATE Sep 79
NOTE 29p.: Paper presented at the Research Conference on Educational Environments and the Undergraduate Women (Wellesley, MA, September 13-15, 1979) ; Appendix may be marginally legible

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests: Aptitude Tests: *College Entrance Examinations: Complexity Level: Elementary Secondary Education: *Females: Higher Education: Item Analysis: Measurement Techniques: Psychometrics: Research Reviews (Publications): Sex Differences: *Sex Discrimination: Student Testing: *Test Bias: Test Construction: Test Interpretation: Test Items: Test Validity

ABSTRACT

Empirical studies are reviewed of sex bias and psychometric bias in aptitude and achievement tests. It is concluded that: (1) of four studies of sex bias, all found that the language used in the tests favored males, but, at the elementary-secondary level, sex bias has been decreasing: (2) of 15 analyses of psychometric bias, all found an item-by-sex interaction effect on performance: (3) of two studies which examined both sex and psychometric bias, one found a relationship between sex bias and sex differences in performance: (4) no evidence that changes in sex bias relate to changes in psychometric bias was noted: and (5) most recent studies of sex bias divert attention from the relative performance of males and females to the item difficulty within sex group: but do not point out that males outperform females overall on mathematics and science tests and that there are basically no items (by content or item type,) on which females outperform males. However, the author cautions interpreting this to imply innate sex differences in mathematical or scientific abilities because it has been determined that, given equal training, such differences disappear.
(Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181077

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

DEC 1979
RECEIVED
ERIC/TM

DRAFT (c

SEX BIAS IN EDUCATIONAL APTITUDE AND ACHIEVEMENT TESTS: WHAT DO WE KNOW?

Marlaine E. Lockheed

Division of Educational Research and Evaluation
Educational Testing Service

"The shop seemed to be full of all
manner of curious things--but the
oddest part of it all was that,
whenever she looked hard at any
shelf, to make out exactly what it
had on it, that particular shelf
was always quite empty, though the
others round it were crowded as full
as they could hold.
"'Things flow about so here!' she
said at last in a plaintive tone..."

- Through the Looking-Glass

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Marlaine LocKheed
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM010038

A paper prepared for the Research Conference on Educational Environments and the Undergraduate Women, Wellesley College, September 13-15, 1979.
Not to be quoted or cited without written permission of the author.

SEX BIAS IN EDUCATIONAL APTITUDE AND ACHIEVEMENT TESTS: WHAT DO WE KNOW?

Marlaine E. Lockheed

Division of Educational Research and Evaluation
Educational Testing Service

The inquiry that prompted this paper was "Is there sex bias in ability tests used in higher education?" Although this question appears to be a simple one, the answer to it depends upon answering the questions "What is sex bias?" and "What are ability tests?" As Williams (1979) notes: "while ability testing has found itself in the arena of public debate, there is often a lack of common understanding of the basic terms and definitions necessary for constructive consideration of this issue" (p. 2). This observation is particularly relevant to the issue of sex bias in tests, and therefore, let me begin this paper with a brief summary of the types and uses of ability tests and of the definitions of sex bias used in the literature.

Definition of terms

Tests. First of all, what are ability tests? In general, there are two kinds of ability tests: aptitude tests and achievement tests. "Aptitude tests measure reasoning ability not directly related to any specific curriculum and assume a common knowledge base" (Williams, 1979, p. 2). The Scholastic Aptitude Test, for example, is a "multiple-choice test made up of separately timed and scored verbal and mathematical sections. Verbal questions measure [the student's] ability to understand what [he or she reads] and the extent of [his or her] vocabulary. Mathematical questions measure [his or her] ability to solve problems involving arithmetic reasoning, algebra and geometry" (Taking the SAT, 1978, p. 1).

In the past, aptitude tests have been used in higher education to screen applicants for college, graduate and professional schools. With declining college enrollments, the importance of aptitude tests for screening may decrease in the future.

"Achievement tests measure developed ability in a particular subject or discipline... where it is recognized that levels of knowledge or competence may differ." (Williams, 1979, p. 2). Achievement tests have been used for certification and evaluation. Increased competition for jobs may generate a widespread demand for such measures. Achievement tests have also been used for admission or selection when a given level or knowledge is a prerequisite for participation in a particular course. For example, a knowledge of algebra may be needed before a student can take calculus, or a basic understanding of Spanish is necessary before a student can enroll in an intermediate Spanish course.

Bias. Bias may be defined from both social and psychometric points of view. The psychometric literature on test bias is extensive and cannot be reviewed here (see Linn, 1973, 1976; and Cronbach, 1975, for examples). Generally speaking, "bias in measurement occurs when characteristics of the measure, the measurement process, or the interpretation of the results of the measurement lead to inaccurate inferences about the knowledge, skills, or other attributes of an individual or a group" (Anderson, Ball & Murphy, 1975, p. 43). In fact "bias is really a matter of invalidity; by definition, valid measurement cannot be biased" (Anderson et al, 1975, p. 43). There are two general methods for determining validity: 1) inspecting the test or test item(s), and 2) correlating the test-takers performance on the test or test item with other measures. The definition of bias from a social perspective relates

to aspects of a test, test process, or interpretation that may be inherently unfair--whether or not these aspects have any effect on the test taker's performance.

Sex bias. The term "sex bias in tests" has been applied both to the test content and to its use; sex bias in the measurement process and sex bias in interpretation have largely been ignored in empirical studies of bias. Two types of sex bias in content have been identified:

1. Gender bias, including sex balance, sexism and sex stereotyping in tests and items (Lockheed, 1974).
2. Psychometric bias, including test specification factors that may influence sex differences in performance, such as item types, cognitive skills required of the test takers, the sex-relevant content, and test balancing (Dwyer, 1976).

In this paper, I will review empirical studies of gender bias and psychometric bias in aptitude and achievement tests. Particular attention will be paid to tests used in higher education.

Summary of empirical studies of gender bias

Gender bias is an aspect of a test that informally acknowledges, reinforces and communicates cultural beliefs about sex differences in abilities, interests or personal charactistics. This type of bias has been referred to as "facial bias" (Cole, 1978), "content bias" (Dwyer, 1976), and simply "sex bias" (Tittle, 1974)--terms subject to multiple interpretations. Gender bias is the opposite of sex fairness, which calls for an equal representation of males and females and of their interests or likes. Gender balance in representation implies that test items be balanced with respect to male and female references, that the status of males and females within items be balanced or equal, that socially determined interests or talents of males and females be equally represented, and that sex

stereotypes be absent(Lockheed, 1974). Over the past several years, considerable improvement in gender bias in tests may be observed.

Imbalance in the "gender ratio" within tests was first investigated in 1973 by Tittle, McCarthy, and Steckler (1974), who studied the occurrence of male and female references in nine different series of standardized tests of academic achievement. Overall, twenty-nine separate tests that were published between 1964-1970 and that were used for elementary and secondary school assessment were examined. Within each test, all references to males and females were counted. The generic use of the word "man" and "he" was counted separately. Of the tests scored, all but one contained a higher number of male references than female reference. The male-female ratio ranged from 14.11 to slightly less than 0.86. Only eight of the twenty-nine test batteries examined had less than a 2.0 ratio of male references to female references. The older the age group for whom the test was written, the higher the male/female ratio.

At about the same time, Lockheed (1973) conducted a similar investigation of male and female references, by item, in eight major college and graduate school entrance examinations. Gender imbalance was found within these tests. Items were coded according to whether they contained no sex reference, male only sex reference, female only sex reference or both male and female sex reference. The ratio of male only items to female only items range from 16.0 to 2.0. Only twenty-two of the 1,220 items coded contained references to both men and women. Seventy-five percent of all the items contained no sex reference, but of the remaining twenty-five percent more than four-fifths were "male only items."

Subsequently, Donlon, Ekstrom, Lockheed and Harris (1977) replicated the Tittle et al (1974) study, using four nationally administered elementary and secondary tests published between 1965 and 1974. The range of male-female ratios for references across tests and subtests was 26.5 to 1.57.

By 1978, however, a comparison of gender ratios in the 1970 Metropolitan Achievement Test with those in the 1978 edition showed improvement. While in 1970 the range was from 4.11 to 0.86, by 1978 the range was 1.91 to 0.87 (Jensen & Beck, 1978). Relative balance, therefore, was found in this particular test by 1978.

In tests used in higher education, Strassberg-Rosenberg and Donlon (1975) also examined gender ratios in the 1974 version of the Scholastic Aptitude Test-Verbal. The gender ratio was computed as 1.75. This represents a considerably more sex balanced test than the 1970 version Lockheed reviewed in 1973.

Gender bias is also indicated by sex differences in how the male and female actors are portrayed within items. In their study of sex-role stereotyping in achievement tests, Tittle et al (1974) reported that "women are portrayed almost exclusively as homemakers or in the pursuit of hobbies" (p. 22). Furthermore they reported that "some items imply that the majority of professions are closed to women" (Tittle, et al, 1974, p. 22).

In the investigation of the college and graduate school admissions tests that Lockheed conducted in 1973, the relative status of males and females within any items was coded. Of the twenty-two items (out of 1,220) which contained both a male and a female actor, ten items portrayed the men and women as equal in status; the remaining twelve items showed the men as being higher status than the women. No item on any of the eight tests portrayed a woman in a higher status position than a man, for example, as a female principal with a male teacher or a female lawyer with a male client. Typically females were referred to as mothers, teachers,

secretaries or wives. Males were referred to as lawyers, managers, principals superintendents, doctors or other professionals.

In the 1977 study conducted by Donlon et al, male roles, females roles and neutral roles were coded separately. The identification of particular roles as female, male or neutral was decided by the percentages of females and males found actively engaged in that role as documented by Occupational Characteristics, 1970 Census of Population and other sources. In the tests and subtests we studied, the average number of "male roles" per item ranged from 2.14 to 0.0; female roles from 2.22 to 0.08. Regrettably, although the sex of the role taker was recorded, these data were not analyzed, and hence no information on this form of stereotypy is available for these tests.

Gender bias is also indicated by traditional or stereotyped male or female interests or skills represented in the item context. For example, a mathematical operations question may be posed in the context of counting marbles or counting dolls. Few studies of sex bias in tests have addressed this issue directly, although this general area of concern, which may be considered the cultural relevance of a test, has been covered widely as it relates to the issue of minority representation on tests (Quirk & Medley, 1972; Linn, 1973).

Coffman's (1961) early study of sex bias in the verbal section of the SAT did address this issue. Judgments were made concerning which of the sixty test items would favor men, which would favor women and which would be neutral. As Coffman (1961) noted "of the seven items judged to be easier for men, one involved mechanical knowledge ..., one involved knowledge of the vocabulary of business, and five involved understanding of science .... Of the ten items judged to be easier for women, nine involved words which might be categorized as describing personal

feelings or personality characteristics.... The other item judged to be easier for women was a sentence completion item dealing with...'the arts'" (p. 123).

No other empirical study to my knowledge has, a priori, evaluated items for such stereotypy in interest or skills. Instead, most empirical studies observe sex differences in performance and then review the content of selected items where the differences are most acute. While this procedure may accurately describe the content of items showing sex differentiated performance, it does not permit an independent description of the test context.

Summary of empirical studies of psychometric sex bias

Psychometric bias, as noted previously, is the obverse of validity. As Cole (1978) notes, "the notion of test bias appearing most frequently in the literature of the last decade is the notion that a test is biased if it measures different things in different groups" (p. 4). The source of such bias may come from the measure, the measurement process and the interpretation of the results. Generally speaking, the question of bias is raised whenever the bivariate distributions of test scores differ from one group to another.

Bias may be considered to fall under three rubrics corresponding to the three categories of validity identified by the APA Standards for Educational and Psychological Tests (1974). These three categories of psychometric validity are: construct validity, content validity and criterion-related validity. These terms have been defined as follows: "Construct validity refers to the degree to which scores on a measure permit inferences about underlying traits.

...Content validity usually refers to the degree to which a measure captures a program objective (or objectives). ...Criterion-related validity refers to the degree to which scores on a measure relate to scores on an external criterion" (Anderson et al, 1975, pages 459, 460, & 461).

Bias may be assessed in total tests, subtests and items. Much has been written about the assessment of psychometric bias. In general, content validity is measured by observing the items and observing how well they sample representatively some domain of tasks, while the measurement of construct validity and criterion-related validity depends upon correlations; such bias, therefore, may be present when the pattern of correlations differs between groups.

Total test bias. Empirical studies of sex bias in aptitude tests used for college admissions have focused on the ability of the tests to predict academic performance. Sex differences in the pattern of correlations between tests and grades have been observed.

For example, Seashore (1962) summarized several studies which conlude that women's grades are more accurately predicted by tests than are men's. These conclusions are reiterated by Cole (1973), who analyzed data from students enrolled in nineteen American coeducational colleges and found that standard test and high school grades better predicted women's first term grades than men's first term grades.

The main rationale for examining the predictive validity of a test, however, is for selection purposes. Selection is itself subject to bias. Several models of selection bias have been identified and applied to male-female selection.

The models of selection bias and evidence of fairness using each are:

1. The traditional regression model, where test bias refers to either overprediction or underprediction of criterion measures for different populations using a single regression equation. This method has been found to be biased against women (ACT, 1973; Wild & Dwyer, 1978).
2. The subjective regression model, in which a constant is added to minority group scores to increase the probability of their selection according to the regression model (Darlington, 1971).
3. The equal risk model (Einhorn & Bass, 1971) where all persons who have the same probability of being successful on a criterion measure are selected. According to this model, the use of either separate or combined male-female regression equations has been found to be fair, but impractical (ACT, 1973).
4. The constant ratio model (Thorndike, 1971) where selection by group is made in proportion to that group's success on the test, was found to be unfair to men (ACT, 1973).
5. The conditional probability model (Cole, 1973), in which the probability of being selected is contingent upon achieving a satisfactory criterion score and is not related to group membership, was also found to be unfair to men (ACT, 1973).
6. The threshold utility model, (Petersen & Novick, 1976) in which "the institution selecting the students must determine the degree of preference, or utilities, for each group. If the utilities are the same for male and female groups, the model is the same as the equal

risk model" (Wild & Dwyer, 1978, p. 4-5). No empirical test of this model has been reported.

Item bias. Bias in a test item is most often indicated by item-statistics, such as p-values, deltas, item-total correlations and, more recently, item characteristic curves (Lord, 1977), one parameter (Rasch) models (Wright, Mead, & Draba, 1976), or a grouped empirical item response curve (Scheuneman, 1975). Of these procedures, the p-values and delta approaches have generally been used in empirical studies of sex bias.

The p-value is the percent of the test takers who pass an item that is attempted. This value may be transformed into $\emptyset$, or 2 arcsin p, or into Δ, which is defined as $4_z + 13$ where z is a normal deviate corresponding to p; in either case, it may be considered an index of item difficulty. If an item is more difficult for one group than another it may or may not be biased; if the correlation between one group's score on an item and some other criterion--usually the total test--is not the same as that of the other group, the item may be biased.

Studies of sex bias in items have taken two approaches:

1. Those that measure gender bias and item difficulty separately and then correlate these measures (the experimental or quasi-experimental approach).
2. Those that measure item difficulty only and then examine the characteristics of items for which sex differences in difficulty emerge (the retrospective approach).

The earliest studies of gender bias on item performance were experimental studies that employed few subjects and studied fewer items. In 1959, Milton, for example, reported a study that examined the relative performance--

of men and women on a mathematics test having ten "feminine" items and ten "masculir " items.

Although males solved more problems than did the females, Milton (1959) noted that "the difference between the scores of men and women was reduced when the problems were stated with content appropriate to the feminine role" (p. 706). These findings were not supported in a subsequent partial replication (Hoffman & Maier, 1966), however. More recently, Graf and Riddell (1972) found that while the "masculinity" or "femininity" of a problem did not affect the subjects' performance on the item, the gender bias of the item did effect speededness for females. Female subjects solved a yardage problem in two-thirds the time it took them to solve a stock market problem.

Three quasi-experimental studies of the effects of gender bias on performance have been identified; however, only two of them have studied nationally administered ability tests for higher education. Examining the college bound students of 1954, Coffman (1961) examined items on the Sat-Verbal aptitude test for gender bias and made predictions as to which items would favor men and which would favor women. The items were judged for traditional interest or skills of men and women; fourteen of the sixteen predictions based on these judgments were confirmed. Of the nine items involving mechanical knowledge, science or business that were judged to be easier for men, eight actually were easier for male test takers. Of the ten items involving personal feelings or personality characteristics that were judged easier for women, nine were actually easier for female test takers.

In a more recent study of college-bound students, Strassberg-Rosenberg & Donlon (1975) found that although the SAT-Verbal test was "generally stereotypical,"

"reference to male or female characters and pronoun usage did not seem to influence either the delta value or bias index (D-value) of the individual verbal items" (Strassberg-Rosenberg & Donlon, 1975, p. 14).

A study of gender bias effects on sex differences in performance for elementary and secondary achievement tests, did report a relationship between content and performance. Girls in the fifth, eighth and ninth grades performed better than boys on test items in which female characters equalled or outnumbered male characters (Donlon et al, 1977).

Retrospective studies of gender bias and sex bias have occurred frequently in the literature and dominate the body of current research. Retrospective studies suffer from two drawbacks. First, the indicator of psychometric bias most often used is the delta plot or D-value (Angoff & Ford, 1973). This is a measure "based on item performance relative to other items, not on absolute differences between the sexes" (Strassberg-Rosenberg & Donlon, 1975, p. 20). That is, items identified by this statistic as "favoring" females may nonetheless be succeeded on more often than males. Second, the logic of the study is faulty, since these studies focus first on the dependent variable and then examine extreme cases. In general, the extreme cases are few, less than 10% of the total test. They are, moreover, arbitrarily determined, and may represent no more than slight random fluctuations.

The first of such retrospective studies was conducted by Donlon (1973) who examined for sex differences by item the scores of the 103,275 persons who took the Scholastic Aptitude Test in May 1964. Items for which the women's performance was statistically superior to the men's, and items for which the men's performance was superior to the women's were located and examined for sex related bias.

Of the ninety verbal items on the test, eight items favored men and eleven favored women. Seven of the eight items favoring males were coded as having scientific or "practical affairs" content; eight of the eleven items favoring females were coded as having human relations, humanities or aesthetic-philosophical content. Of the sixty mathematical items on the test, twenty-nine favored men and none favored women. The items that were the relatively easiest for men dealt with "real world" topics, while the items that were relatively easiest for women dealt with algebra. This is not to say, however, that women performed better on albegra items than men did.

Two years later, Strassberg-Rosenberg and Donlon (1975) reexamined the 1974 SAT for sex bias. The measure used was delta plots, in which item difficulties (deltas) for males and females were plotted on a single graph, and outliers identified from the major axis of the elliptical plot for the highly correlated delta values. The use of this statistic makes it difficult to compare the p-values for males and females, even though delta is based upon a transformation of p. This statistic, moreover, leads to ambiguities in the interpretation of findings.

For the verbal part of the SAT, Strassberg-Rosenberg and Donlon (1975) identified twelve out of ninety items as "outliers"--items for which the difficulty index (delta) for males was different than for females. Of these, nine were relatively easiest for males and three were relatively easiest for females. The relatively easiest items for males were on topics related to science and practical affairs; the relatively easiest items for females were on topics related to asthetics and philosophy. Inspection of the delta plots for this test also shows that not only were these items relatively easiest for each sex, they were also absolutely easier in comparison with the other sex. Thus, although overall the SAT-Verbal exhibited no sex differences, if the thirty science and practical affairs items had been matched in number to equal the twenty-four aesthetic-philosophical and human

relations items, women's SAT-Verbal scores might have been higher than men's scores.

The mathematics component of the 1974 SAT covered three domains of mathematics--geometry, algebra, and arithmetic--and included several miscellaneous items. With one exception, men outperformed women on all mathematics items. Nevertheless five outliers "favoring" men and two outliers "favoring" women were identified. Outliers relatively easiest for men covered geometry and arithmetic, while outliers relatively easiest for women covered algebra (Strassberg-Rosenberg & Donlon, 1975).

A third retrospective study of sex bias in the SAT was conducted by Stern (1978), who examined the performance of 1,000 black males, 1,000 black females, 1,000 white males and 1,000 white females who were matched as to verbal ability. No sex difference on the verbal items of the SAT was found for the black sample, whereas white females did better on verbal items except at the upper delta scale where the lines either converged or slightly crossed. By comparison, black males were slightly more able on mathematics and white males consistently had less difficulty with mathematics across the entire delta scale.

Outliers favoring males and females were identified in this study as well. For the verbal test the six items relatively easiest for black females included science, aesthetics, and reading passages; the three items relatively easist for black males included science and practical affairs. The ten items relatively easiest for white females included science, aesthetics, human relationships and reading passages. Because the delta plots for both white students and black students fell near the 45° line, it is reasonable to presume that the items that were relatively easier for each sex were also absolutely easier.

For the math test, however, it is not possible to assume this, as the delta plots indicate clear sex differences in performance in favor of males for both black and white groups; in particular, for white students, females outperformed males on only two of the sixty items. Geometry items were the most difficult mathematics item for white females.

Stern (1978) did not confuse relative difficulty within sex with relative difficulty betwen sex. Studies by Hicks, Donlon and Wallmark (1976) and by Wheeler and Harris (1979) used language that failed to clarify this distinction, however.

Hicks et al (1976) studied the performance of 1,720 white males and 1,735 white females who took the GRE aptitude test in 1974. Overall, few sex differences in either vocabulary or reading were found; males outperformed females on the mathematics section. The delta plots for the mathematics section showed only one item out of fifty-five on which females outperformed males. Nevertheless the authors note that, of the nine mathematical outliers, "five outliers were relatively favorable to females, four to males" (p. 13) and conclude that "the test is not overly unfair to one of the sexes" (Hicks et al, 1976, p. 15). The reader who is unfamiliar with the interpretation of the D-value might erroneously conclude that there were, on this test, five items on which women outperformed men, four items on which men outperformed women, and on the rest of the items the two sexes performed about equally. This would be an inaccurate interpretation.

A similarly misleading interpretation could be drawn from the conclusion of an otherwise exemplary recent study of the ATP Physics test (Wheeler & Harris, 1979). This study examined the performance of 710 females and 802 males who took the ATP Physics test in 1978. From test statistics and delta plots it is clear that on seventy-four of the seventy-five items of the test, males outperformed

females. Yet, Table 9 of this study reported the content of seven outlier items to be "relatively easier for males" and seven outlier items to be "relatively easier for females." Furthemore, the summary notes that "none of the five content areas represented on the test... stood out as being relatively easier or relatively harder for males than for females" (Wheeler & Harris, 1979, p.30). The intended meaning of this statement is that neither men nor women found any content area easier or harder relative to other content areas. It is not difficult to see, however, that this summary could be interpreted to mean that no content area was easier or harder for males than for females.

This study of the ATP was exemplary because it was unique in utilizing several indicators of sex bias, including item difficulty, item discrimination, speededness, omits, and error analysis. The latter three analyses provide some evidence for bias, suggestive that some aspect of the measurement process may inhibit female performance. That is, women were less likely to complete the test, and were twice as likely to omit items, particularly the first ten items. Error analysis also revealed sex differences that indicated the girls did not understand basic concepts while the boys "knew too much." Both test anxiety and lack of physics-related experience may have contributed to the observed differences.

Conclusion

This paper has reviewed the empirical evidence for asserting that ability tests used in higher education exhibit sex bias. Two types of sex bias -- gender bias and psychometric bias -- have been identified. Of the four studies that examined gender bias, all found that the language of the tests favored males over females; from studies of gender bias in elementary and secondary level tests,

it is clear that gender bias has decreased over time. Of the fifteen analyses for psychometric bias, all found an item by sex interaction effect on performance; this effect was noted for 142 (16.1%) of the 825 items analyzed over all studies. Of the two studies that examined both gender and psychometric bias, one found a relationship between gender bias and sex differences in performance; the other did not. No evidence that changes in gender bias relate to changes in psychometric bias has been found.

The most recent studies of sex bias, by emphasizing outlier items from delta-plots, divert attention from the relative performance of males and females to the item difficulty within sex group. This approach is generally misleading and fails to point out the most obvious findings of these studies: that men outperform women overall on mathematics and science tests and that there are virtually no items -- by content or item type — on which women outperform men.

This finding should not be interpreted as implying innate sex differences in mathematics and science abilities. Such differences disappear when men and women with equivalent training and experience have been compared (deWolf, 1977; Wheeler & Harris, 1979). In all probability bias in the educational experience of the test takers rather than bias in the test itself accounts for sex differences in performance in these areas.

References

American College Testing Program. Bias in the uses of educational assessment. In Assessing students on the way to college (Volume 1). Iowa City, Iowa: The American College Testing Program, 1973.

American Psychological Association. Standards for educational and psychological tests. APA, Washington, 1974.

Anderson, S., Ball, S., & Murphy, R. Encyclopedia of educational evaluation. New York: Jossey-Bass, 1975.

Angoff, W. H., & Ford, S. F. Item-race interaction on, a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Coffman, W. E. Sex differences in responses to items in an aptitude test. Eighteenth Yearbook, National Council on Measurement in Education, 1961, 117-124.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10(2), 237-255.

Cole, N. S. Approaches to examining bias in achievement test items. Association for Measurement and Evaluation in Guidance Newsnotes, Spring 1978, 13(3), 4-8.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Darlington, R. B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

deWolf, V. High school mathematics preparation and sex differences in quantitative abilities (EAC Report 77-17). Seattle, Wash.: Educational Assessment Center, 1977.

Donlon, T. F. Content factors in sex differences on test questions (RM 73-28). Princeton, N.J.: Educational Testing Service, 1973.

Donlon, T. F., Ekstrom, R. B., Lockheed, M. E., & Harris, A. Performance consequences of sex bias in the content of major achievement batteries (PR 77-11). Princeton, N.J.: Educational Testing Service, 1977.

Dwyer, C. A. Test content in mathematics and science: The consideration of sex. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Einhor, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75(4), 261-269.

Graf, R. G., & Riddele, J. C. Sex differences in problem-solving as a function of problem context. The Journal of Educational Research, 1972, 65, 451-452.

Hicks, M. M., Donlon, T. F., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. National Council on Measurement in Education, 1976.

Hoffman, L. R., & Maier, N. R. F. Social factors influencing problem-solving in women. Journal of Personality and Social Psychology, 1966, 4, 382-390.

Jensen, M., & Beck, M. D. Gender balance analysis of the metropolitan achievement tests, 1978 Edition. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 28, 1978.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43(2), 139-161.

Linn, R. L. In search of fair selection procedures. Journal of Educational Measurement, 1976, 13(1), 53-58.

Lockheed, M. E. Test bias in ETS tests. Unpublished memorandum, Educational Testing Service, Princeton, N.J., 1973.

Lockheed, M. E. Sex bias in educational testing: A sociologists's perspective (RM 74-13). Princeton, N.J.: Educational Testing Service, 1974.

Lord, F. M. Practical applications of item characteristic curve theory (RM 77-3). Princeton, N.J.: Educational Testing Service, 1977.

Milton, G. A. Sex differences in problem solving as a function of role appropriateness of the problem content. Psychological Reports, 1959, 5, 705-708.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13(1), 3-29.

Quirk, T. J. & Medley, D. M. Race and subject-matter influences on performance on general education items of the National Teacher Examinations (RB 72-43). Princeton N.J.: Educational Testing Service, 1972.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975.

Seashore, H. G. Women are more predictable than men. Journal of Counseling Psychology (Vol. 9), 3, 261-277.

Stern, J. College Board item bias study of the Scholastic Aptitude Test and the Test of Standard Written English (SR 78-56) Form XSA2. Princeton, N.J.: Educational Testing Service, 1978.

Strassberg-Rosenberg, B., & Donlon, T. F. Content influences on sex differences in performance on Aptitude Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 28, 1975.

Taking the SAT. Admissions Testing Program of the College Board, College Entrance Examination Board, New York, 1978.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8(2), 63-70.

Tittle, C. K. Sex bias in educational measurement: Fact or fiction? Measurement and Evaluation in Guidance, 1974, 6(4), 219-226.

Tittle, C. K., McCarthy, K., & Steckler, J. F. Women and educational testing. Princeton, N.J.: Educational Testing Service, 1974.

Wheeler, P., & Harris, A. Performance differences between males and females on the ATP Physics Test. Berkeley, California: Educational Testing Service, 1979.

Wild, C. L., & Dwyer, C. A. Sex bias in selection (RM 78-1). Princeton, N.J.: Educational Testing Service, 1978.

Williams, E. B. Ability testing and affirmative action: Some theoretical and practical considerations. Paper presented at the annual meeting of the American Psychiatric Association, Chicago, Il., May 14-18, 1979.

Wright, B. D., Mead, R. J., & Draba, R. E. Detecting and correcting test item bias with a logistic response model (Research Memorandum No. 22). Statistical Laboratory, Department of Education, University of Chicago, 1976.

Empirical Studies of Test Characteristics and Sex Differences in Performance

on Higher Education Aptitude and Achievement Tests

Test and Reference	Data Characteristics: Year	Sample	Number of Items	Analyses: Measure of Sex-Related Content	Results	Measure of Sex Diff. in Performance	Results	Characteristics of Items Favoring Males and Females
Scholastic Aptitude Test-Verbal (SAT-V)								
Coffman (1961)	1954	370 female 370 male	60	Judgments about item difficulty	7 items judged easier for males; 10 judged easier for females; 43 judged neutral	$\phi > \pm 1.96$; $p < .05$ (adjusted difference in percent passing)	6/7 "male" items easier for males; 8/10 "female" items easier for females	Item content: easiest items for males are mechanical knowledge, science, business; easiest items for females are personal feelings, personality characteristics
Donlon (1973)	1964	47,083 female 55,717 male	90	none	none	$p > \pm .07$ (difference in percent passing)	8 items easier for males; 11 items easier for females	Content of items easiest for males are science or practical affairs; content of items easiest for females are humanities or aesthetics
Strassberg-Rosenberg & Donlon (1975)	1974	1,000 female 1,000 male	90	ratio of male nouns or pronouns to female nouns or pronouns	1.75 male references to each female reference	$\lvert D_i \rvert \geq 1.5$ s.d. of D (relatively easiest items for each sex)	p not provided; 9 items relatively easiest for males; 3 items relatively easiest for females; No relationship between M/F ratio and item difficulty	Easiest items for males are science or history; easiest items for female are human relations and female stereotyped material
Stern (1978)	1975	Black 1,000 female 1,000 male White 1,000 female 1,000 male	80	none	none	$\lvert D_i \rvert \geq 1.25$ s.d. of D (relatively easiest items for each sex)	Black: 3 items relatively easiest for males; 6 items relatively easiest for females White: 7 items relatively easiest for males; 10 items relatively easiest for females	For blacks, no sex differences in content difficulty; for whites, content of items easiest for females are humanities and aesthetics; conte of items easiest for males are scien and practical affairs
Scholastic Aptitude Test-Mathematics (SAT-M)								
Donlon (1973)	1964	47,082 female 55,717 male total population	60	none	none	$p > \pm .07$	29 items easier for males; 0 items easier for females	Male item content: all subjects; easiest items for males are "real world" subjects; easiest items for females are algebra items

22.

Empirical Studies of Test Characteristics and Sex Differences in Performance

on Higher Education Aptitude and Achievement Tests

Test and Reference	Data Characteristics			Analyses				
	Year	Sample	Number of Items	Measure of Sex-Related Content	Results	Measure of Sex Diff. in Performance	Results	Characteristics of Items Favoring Males and Females
Strassberg-Rosenberg & Donlon (1975)	1974	1,000 female 1,000 male random sample	60	ratio of male nouns or pronouns to female nouns or pronouns	in 12 "real world" items, the M/F ratio = 2.25	$\|D_i\| \geq 1.5$ s.d. of D	59 items easier for males; 1 item easier for females On total test: 3 items relatively easiest for males; 3 items relatively easiest for females. On regular math: 2 items relatively easiest for male; 3 items relatively easiest for females; On data sufficiency 2 items favored female	Male item content: all subjects; easiest items for males are geometry and arithmetic Female item content: no subjects; easiest items for females are algebra items
Stern (1978)	1975	Black 1,000 female 1,000 male White 1,000 female 1,000 male matched on verbal scores	60	none	none	$\|D_i\| \geq 1.25$ s.d. of D	Black: 5 items relatively easiest for males; 3 items relatively easiest for females White: 4 items relatively easiest for males; 2 items relatively easiest for females	For blacks, no sex difference in content difficulties For whites, no particular female content; easiest items for males are arithmetic and geometry
Test of Standard Written English								
Stern (1978)	1975	Black 1,000 female 1,000 male White 1,000 female 1,000 male matched on verbal scores	50	none	none	$\|D_i\| \geq 1.25$ s.d. of D	Black: 1 item relatively easiest for males; 3 items relatively easiest for females White: 2 items relatively easiest for males; 3 items relatively easiest for females	For blacks, no sex difference in content difficulty For whites: no particular male content in items; easiest items for females are sentence corrections.
Graduate Record Examination-Verbal								
Hicks, Donlon, & Wallmark (1976)	1974	1,735 female 1,720 male	55 vocab 40 reading	none	none	$\|D_i\| \geq 1.5$ s.d. of D	Vocabulary: p not provided; 2 items relatively easiest for male; 3 items relatively easiest for female Reading: p not provided; 3 items relatively easiest for males; 3 items relatively easiest for females	Item content: Easiest items for males are practical affairs and science; easiest items for females are human relations. practical affairs Item content: easiest items for males are physical science; easiest items for females are humanities and argumentative

Empirical Studies of Test Characteristics and Sex Differences in Performance

on Higher Education Aptitude and Achievement Tests

Test and Reference	Data Characteristics			Analyses				
	Year	Sample	Number of Items	Measure of Sex-Related Content	Results	Measure of Sex Diff. in Performance	Results	Characteristics of Items Favoring Males and Females
Graduate Record Examination-Math Hicks, Donlon, & Wallmark (1976)	1974	1,735 female 1,720 male	55	Male/female values for referents in all items not provided	male/female values for referents in all items not provided		Math: p's not provided; 4 items relatively easiest for males; 5 items relatively easiest for females	Item content: easiest items for males are data interpretation, geometry; easiest items for females are algebraic unknowns
ATP Physics Achievement Test Wheeler and Harris (1979)	1978	710 female 802 male total population of 710 females matched sample of 802 males	75	none	none	$D \geq 1.25$ s.d. of D	On total test: 74 items easiest for males; 1 item easiest for females; 7 items relatively easiest for males; 7 items relatively easiest for females	Item content: easiest items for males are pictorial, verbal, and electricity and magnetism and knowledge and application Item content: easiest items for females are graphic, numerical, and symbolic and comprehension and application
		710 female 802 male		none	none	percent omit	23 items omitted by twice as high a proportion of women as men (included 6 items identified as relatively easiest for males); 3 items omitted by higher proportion of males (included 2 items identified as relatively easiest for females)	19 omitted items are pictorial and verbal format; 20 of the 23 omitted items are application and knowledge 2 of the 3 omitted items are mechanics with graphic-numerical format; 1 item is numerical format in heat and kinetic theory

24.